A SYSTEM AND METHOD FOR WORKLOAD-AWARE REQUEST DISTRIBUTION IN CLUSTER-BASED NETWORK SERVERS

ABSTRACT OF THE DISCLOSURE

A method and system for workload-aware request in cluster-based network servers. The present invention provides a web server cluster having a plurality of nodes wherein each node comprises a distributor component, a dispatcher component and a server component. In another embodiment, the present provides a method for managing request distribution to a set of files stored on a web server cluster. A request for a file is received at a first node of a plurality of nodes, each node comprising a distributor component, a dispatcher component and a server component. If the request is for a core file, the request is processed at the first node (e.g., processed locally). If the request is for a partitioned file, it is determined whether the request is assigned to be processed locally at the first node or at another node (e.g., processed remotely). If the request is for neither a core file nor a partitioned file, the request is processed at the first node. In one embodiment, the present invention provides a method for identifying a set of frequently accessed files on a server cluster comprising a number of nodes. Embodiments of the present invention operate to maximize the number of requests served from the total cluster memory of a web server cluster and to minimize the forwarding overhead and disk access overhead by identifying the subset of core files to be processed at any node and by identifying the subset of partitioned files to be processed by different nodes in the cluster.